Proc. Natl. Acad. Sci. USA
Vol. 92, pp. 11057–11060, November 1995
Plant Biology

The gene distribution of the maize genome

(plants/Gramineae/isochores/chromosomes)

NICOLAS CARELS, ABDELALI BARAKAT, AND GIORGIO BERNARDI*

Laboratoire de Génétique Moléculaire, Institut Jacques Monod, 2, Place Jussieu, 75005 Paris, France

Communicated by Gary Felsenfeld, National Institutes of Health, Bethesda, MD, August 11, 1995

ABSTRACT Previous investigations from our laboratory showed that the genomes of plants, like those of vertebrates, are mosaics of isochores, i.e., of very long DNA segments that are compositionally homogeneous and that can be subdivided into a small number of families characterized by different GC levels (GC is the mole fraction of guanine + cytosine). Compositional DNA fractions corresponding to different isochore families were used to investigate, by hybridization with appropriate probes, the gene distribution in vertebrate genomes. Here we report such a study on the genome of a plant, maize. The gene distribution that we found is most striking, in that almost all genes are present in isochores covering an extremely narrow (1–2%) GC range and only representing 10–20% of the genome. This gene distribution, which seems to characterize other Gramineae as well, is remarkably different from the gene distribution previously found in vertebrate genomes.

The genomes of vertebrates are made up of isochores (1, 2), long (>300 kb) DNA segments that are homogeneous in base composition and that can be subdivided into a small number of families covering a compositional range, which is broad in warm-blooded vertebrates and narrow in cold-blooded vertebrates. The compositional distribution of isochores is reflected in the DNA fragments (≈100 kb long) that derive from them as a result of the enzymatic and physical degradation accompanying DNA preparation and that can be fractionated by centrifugation in Cs_2SO_4 density gradients in the presence of sequence-specific ligands. The major interest of this approach is that it has allowed the definition of genome properties such as the compositional patterns (e.g., the compositional distributions of DNA fragments; see *Results*), the compositional correlations between coding and flanking noncoding sequences, and above all, the distribution of genes in the genome (1–4).

The same approach has demonstrated that the genomes of angiosperms also consist of isochores (5–7) and that the compositional patterns of Gramineae are different from those of other monocotyledons (monocots) and of the dicotyledons (dicots) analyzed. At the DNA level, the compositional patterns of the genomes of Gramineae have a higher average mole fraction of G + C (GC) and are generally wider than those of the other angiosperms tested (5). At the level of coding sequences, compositional patterns not only are wider but also are characterized by an abundance of sequences with a high GC level in third codon position (GC_3) and by biased codon usage (6). As far as the overall distribution of genes in the genome is concerned, to our knowledge, no information is available for plants, even if some features related to such a distribution have been studied by other authors (ref. 7; see *Discussion*). We have investigated this problem for the maize genome and have discovered that almost all maize genes are located in isochores covering an extremely narrow (1–2%) GC range, representing only 10–20% of the genome. This striking gene distribution, which seems to be shared by other Gramineae, is very different from those previously found in vertebrate genomes. We discuss its practical implications.

MATERIALS AND METHODS

DNA Preparations. Maize DNA was obtained from nuclei prepared as described (8) and derived from leaves of cv. F7×F2 (Institut National de la Recherche Agronomique, Versailles, France). The molecular size of DNA fragments ranged from 50 to 150 kb.

DNA Fractionation. Fractionation was performed by preparative centrifugation of maize DNA in a Cs_2SO_4 density gradient containing an AT-specific ligand, 3,6-bis(acetatomercurimethyl)-1,4-dioxane (BAMD) (9), as described in ref. 5, except that the ligand/nucleotide molar ratio (r_f) used was 0.10. Fractions from a single preparative centrifugation were investigated by analytical ultracentrifugation in CsCl density gradient as described (5, 10), and their relative amounts in total DNA were assessed. The analytical CsCl profiles of the fractions indicated a very high molecular weight, as expected from previous extensive evidence for a lack of degradation of DNA submitted to the fractionation procedure used.

Maize Gene Data. Data were from GenBank (release 81.0, March 1994). Coding sequences of multigene families that had the same GC levels in first, second, and third codon positions were assumed to correspond to identical genes and were counted only once.

Restriction Endonuclease Digestion, Electrophoresis, and DNA Filter Hybridization. Aliquots of maize DNA from the Cs_2SO_4/BAMD fractions, proportional to the relative amounts of DNA present in the fractions and corresponding to 20 μg of total DNA, were digested with the restriction endonucleases used by the authors who originally sequenced the genes under analysis (see Table 1). This approach allowed the unambiguous identification of the sequenced genes (primary structures being needed to obtain GC_3 values) after electrophoresis of the digests on 0.8% agarose gels in TAE buffer (0.04 M Tris acetate/0.001 M EDTA, pH 8.0), alkaline transfer onto nylon membranes, and radioactive probe hybridization (11). In some cases, expressed sequence tags were amplified by PCR, radioactively labeled, and used as probes. Again the upper size of restriction fragments, as detected by ethidium bromide staining, and their yield (for example, see Fig. 2) indicate high molecular weights of the fractionated DNA.

RESULTS

The human and maize genomes, two genomes of about the same haploid size (3×10^9 bp), are similar in their compositional patterns. In both, the CsCl profiles (Fig. 1 *A* and *B*) are broad and cover a similar GC range (30% to 60–70%). The GC_3 levels of coding sequences (Fig. 1 *A′* and *B′*) are also

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

Abbreviations: BAMD, 3,6-bis(acetatomercurimethyl)-1,4-dioxane; GC, mole fraction of guanine + cytosine; GC_3, GC level in the third codon position.

*To whom reprint requests should be addressed.

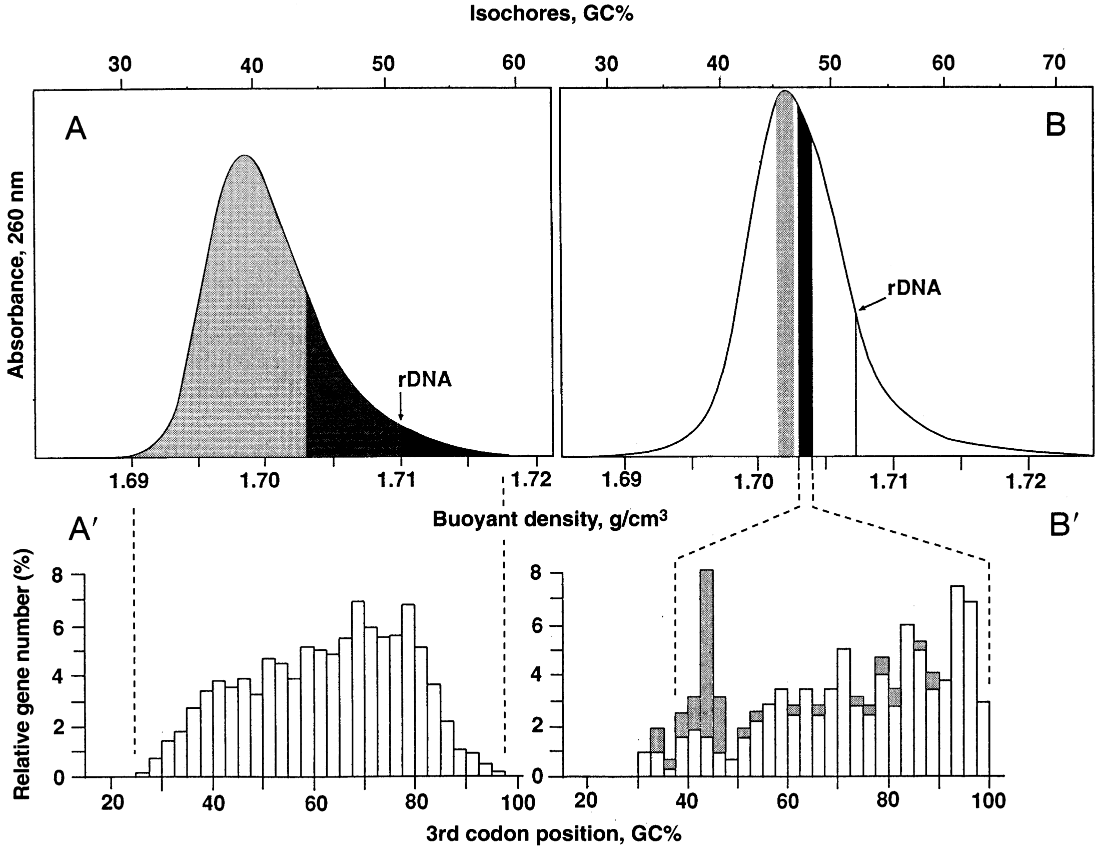

FIG. 1. CsCl profile of human DNA (*A*) is compared with that of maize DNA (*B*). Both profiles were obtained by centrifuging high molecular weight (50–150 kb) DNA preparations in a CsCl analytical density gradient. The GC scale of the maize DNA profile was obtained from HPLC nucleoside analyses (12) of fractions obtained by centrifuging DNA in preparative Cs_2SO_4 density gradients in the presence of BAMD (9). The slightly different relationship of the GC scale to the buoyant density scale in *A* and *B* is due to the fact that the high methylation level of maize DNA (12) causes a shift of its CsCl profile toward lower buoyant density values. For human DNA, solid, darkly shaded, and lightly shaded areas correspond to high, medium, and low gene concentrations in the H3, H1+H2, and L1+L2 (namely the GC-poor, GC-rich, and very GC-rich) isochore families, respectively (for quantitative data, see refs. 3 and 4). For maize, the solid area corresponds to the gene space; zein genes occupy, however, not only the gene space but also DNA compartments corresponding to the lightly shaded area to its left; the open areas represent genome regions where no protein-encoding genes were detected. Ribosomal genes were centered at the buoyant densities marked rDNA. The relative number of genes from human (*A*′) and maize (*B*′) are plotted against GC_3 of the corresponding coding sequences. The histogram of 2762 human genes is from GenBank. Maize data are from GenBank and concern 265 genes plus 41 zein genes (shown as shaded bars). The vertical broken lines correlate the gene distributions (lower frames) with the DNA distributions (upper frames). In maize, such vertical lines only bracket the GC_3 range of the genes tested (see Table 1).

similar, in that they show essentially the same very broad range (25–30% to 97–100%) and are characterized by greater gene frequencies at higher GC_3 levels (neglecting a GC-poor subfamily of zein genes; see below).

In sharp contrast, gene distributions are very different in the two genomes. In the human genome (Fig. 1*A*), genes are present in all isochores, which cover the whole GC range, 30%. Gene concentration is low and constant over the GC-poor isochores of the L1 + L2 families that form two-thirds of the genome, increases in the H1 + H2 isochores of intermediate GC levels, and reaches the highest concentration (at least 20 times higher than in GC-poor isochores) in the GC-richest family of H3 isochores (1–4), which was called the "genome core."

In the maize genome (Table 1, Fig. 2, and refs. 5–7 and 26 contain additional examples of gene localization in compositional fractions), all protein-encoding genes tested, except for some zein genes (see below), were concentrated in a 1% GC interval that represents about 10% of the genome and that we will call the gene space (Fig. 1*B*). The genes probed were chosen at random, covered almost the whole GC_3 range of the maize genes available in gene databanks (Fig. 1*B*′), were evenly distributed in terms of GC_3, and encoded functionally different proteins. It should be noted, however, that in Table 1, the modes of the CsCl peaks of the DNA fractions showing the strongest hybridization intensities were considered to represent the modes of the distribution of the corresponding genes. This approximation may lead, however, to an underestimate of the gene space, which might, in fact, be between 1% and 2% GC and correspond to 10–20% of the genome. As for zein genes, they were found not only in the gene space but also in GC-poorer compartments reaching buoyant densities as low as 1.7013 g/cm³ (ref. 7 and unpublished data), this value probably corresponding to the subfamily of the zein genes that shows a peak at low GC_3, 45% (Fig. 1*B*′). These GC-poorer compartments (Fig. 1*B*′) are characterized by a much lower gene density than the gene space, since as far as we know, they only contain a subset of a large gene family. Finally, the GC-rich ribosomal genes were localized in a GC-rich compartment (1.7070 g/cm³) that corresponded to a satellite peak in the Cs_2SO_4/BAMD preparative density gradient (data not shown).

DISCUSSION

The results presented above deserve several comments.

(*i*) The gene space may be visualized as a single family of isochores, in that the GC range of all the isochores containing protein-encoding genes (except for some of the zein genes) is narrower (1–2% GC) than that exhibited by any of the human

Table 1. Gene location in compositional fractions of maize DNA

Genes		DNA fractions	
Name	GC_3	ρ, g/cm^3	GC
Zeins		From *1.7013*	45.3
		to 1.7030	47.2
Fe. Glu. Synth.	37.2	1.7030	47.2
ANT1	53.6	1.7030	47.2
NDME*	58.2	1.7032	47.4
cdc2 kinase	63.6	1.7030	47.2
ADH-1†	65.8	*1.7030*	47.2
POD*	69.8	1.7032	47.4
Sucrose P S	70.2	1.7032	47.4
Tubulin	70.6	1.7032	47.4
R-S	78.0	1.7032	47.4
Knotted	80.2	1.7032	47.4
Brittle	83.9	1.7034	47.6
PEPcase	84.1	*1.7032*	47.4
Wx	92.7	*1.7032*	47.4
H4	94.2	*1.7032*	47.4
Cab*	94.3	1.7032	47.4
GAPDH*	96.5	1.7032	47.4
rbcs	97.0	*1.7038*	48.1
Chalcone S	98.0	*1.7032*	47.4
T12672*	‡	1.7032	47.4

Fe. Glu. Synth., ferredoxin-dependent glutamate synthase (13); ANT 1, adenine nucleotide translocator (14); NDME, NADP-dependent malic enzyme (15); cdc2 Kinase, *cdc2* protein kinase (16); ADH-1, alcohol dehydrogenase (7); POD, pyruvate orthophosphate dikinase (17); tubulin, α-tubulin (18); sucrose P S, sucrose phosphate synthase (19); R-S, regulatory gene of maize anthocyanin synthesis (20); knotted, knotted (21); brittle, brittle-1 (22); PEPcase, phospho*enol*pyruvate carboxylase (7); Wx, waxy (7); H4, Histone H4 (7); Cab, light-harvesting chlorophyll a/b binding protein (23); GAPDH, glyceraldehyde-3-phosphate dehydrogenase (24); rbcs, ribulose-1,5-biphosphate carboxylase; chalcone S, chalcone synthase (7); T12672, unknown function (C. Baysdorfer, 1993, personal communication to the Maize Genetic Database). GC_3 is the average GC level in third codon positions. The buoyant density indicated is that of the fraction showing the highest hybridization intensity. Values from ref. 7 are in italic type. Other values are from the present work. GC values were calculated from buoyant density values by using the equation $GC(\%) = [1102 \times \rho(g/cm^3)] - 1829.5$, which is derived from the data of ref. 12. Zein genes were detected with the pML1 (25) probe (GC_3 = 46.4%).

*Expressed sequence tag probes (GenBank release 87.0, February 1995).

†The buoyant density value for ADH1 quoted in table 1 of ref. 7, 1.7018 g/cm^3, was wrong. The correct value 1.7030 g/cm^3 can be deduced from figure 2B of ref. 7 and figure 2B of ref. 5.

‡Partial sequence.

isochore families. The fact that introns of maize genes are both scarce and short could contribute to the compactness of the gene space. All other isochores (with the exception of those containing some of the zein genes and ribosomal genes) appear to correspond to gene-empty space. This situation is strikingly different from that found in the human genome, where genes are spread over all isochores, which cover a broad GC range, 30% (see Fig. 1*A*).

(*ii*) Even though the gene sample tested contained only 21 genes, these were chosen at random, covered almost the whole range of GC_3 values of maize genes, and encoded functionally different proteins. In other words, the sample used should be a representative one, and the conclusions drawn from it can be extrapolated to other maize genes. It should be noted that the 21 probes tested detected not only the specific genes from which they were derived but also related genes from the same families (see, for example, Fig. 2). These related genes, for which we do not have sequence information, were also located in the gene space. In other words, many more than 21 genes were located in the gene space studied with the 21 probes.

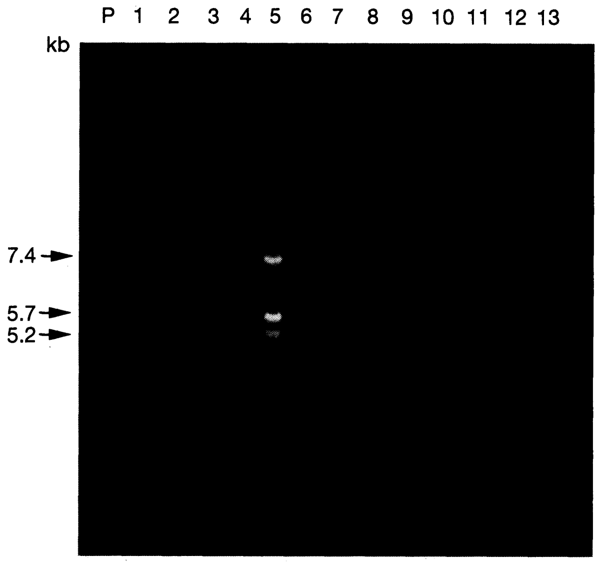

FIG. 2. Southern blot analysis showing the location of adenine nucleotide translocator (*ANT*) genes in compositional fractions from maize degraded by *Eco*RI. Arrows indicate hybridization bands. Lanes: P, the (GC-poorest) pellet fraction; 1–13, fractions. The *ANT* gene probe detected two copies of the gene: the first (bands at 5.7 and 7.4 kb) was mainly localized in fraction 5 and corresponds to the described sequence (ref. 14 and A. Baker, B. Winning, and C. J. Leaver, personal communication); the second (band at 5.2 kb) in fractions 4 and 5 corresponds to another gene from the same family. The buoyant densities of fractions 4, 5, and 6 were 1.7024, 1.7030, and 1.7035 g/cm^3, respectively.

(*iii*) The finding that most maize genes are concentrated in a gene space representing 10–20% of the genome and covering an extremely small GC range indicates that most families of repeated sequences are located outside the gene space. Indeed, repeated sequences form at least 80% of the maize genome (27, 28), as opposed to 30% in the human genome, and cannot physically fit into the gene space, which only represents 10–20% of the genome, even though the latter is known to contain some particular sets of repeated sequences in its intergenic regions (see point *iv* below). In this connection, it is of interest to recall that, based on DNA–DNA reassociation kinetics, unique sequences tend to be associated with middle repetitive sequences, whereas highly repetitive and middle repetitive sequences are intermixed (28) and that repetitive elements were found to be spatially separated from single copy sequences in a 280-kb region surrounding the *ADH-1* locus (29).

(*iv*) The gene space in the maize genome appears to correspond to the only genome compartments in which certain mobile sequences can be transposed. For instance, *Mu* (mutator), *Ac* (activator), and the majority of *Cin4* elements are exclusively located in the same class of isochores as the *ADH-1* gene (30), i.e. within the gene space. Maize transposable elements appear, therefore, to be located in gene-rich transcriptionally active isochores. Interestingly, this is in agreement with previous observations on transcribed proviral sequences that also integrate into transcriptionally active regions of mammalian genomes (31) and suggests a correlation between integration and transcriptional activity.

(*v*) The results presented here for the maize genome should be also largely valid for the genomes of other cereals, as suggested by the existence of large regions of collinearity among these genomes, as revealed by genetic linkage mapping (32); intergeneric sexual hybridization experiments (32); homology between coding sequences (6); similarities of isochore patterns (12); and results obtained for wheat (7). In the latter case, hybridization of all of the gene probes tested (histone H3;

rbcs, ribulose-1,5-biphosphate carboxylase; Cab, chlorophyll-a/b binding protein; α-amylase) on wheat fractions yielded results essentially identical to those obtained on maize fractions (7). Similar conclusions can be drawn from hybridization experiments (unpublished data) using maize gene probes. Moreover, probes for genes of storage proteins (high molecular weight glutenin; α,β-gliadin) hybridized to fractions exhibiting slightly lower buoyant densities than other genes (5, 7), as was the case with zein genes in maize.

(*vi*) The gene space consists of a number of small compositional compartments located on all chromosomes, as judged from the known chromosomal location of maize genes (33). Most of the loci tested in wheat physically map in the distal region of the chromosomes (34), where most of recombination also occurs (34, 35). In contrast, the pericentromeric region of a rye chromosome (36) shows a reduced density in unmethylated *Not* I sites, which are indicative of CpG islands associated with genes (37–39). Interestingly, the high gene concentrations and the high recombination levels detected in telomeric regions of the wheat genome are reminiscent of similar findings on the human genome (40, 41) and may reflect a widespread situation.

In conclusion, the maize genome (and probably the genomes of other Gramineae as well) exhibits a very striking gene distribution with almost all genes present in 10–20% of the genome and in a narrow GC range (1–2%). This finding, which was obtained in an overall analysis of the maize genome, fits with previous findings concerning the extremely large amounts of repetitive sequences (27, 28) and their distribution (28, 29) with the high gene concentration near telomeres (34) and with CpG island maps of genome regions (36). From a practical point of view, the use of DNA fractions corresponding to the gene space should facilitate physical and genetic mapping of the maize genome and of other genomes from Gramineae, as the use of the gene-richest compartments did for the human genome (42).

We thank J. Colasanti (Cold Spring Harbor, N.Y.), D. de Vienne and S. Santoni (Gif-sur-Yvette, France), T. Hase (Osaka, Japan), and B. Winning (Oxford, U.K.) for the gift of maize probes; T. Musket (Columbia, MO) for the gift of maize expressed sequence tags; and G. Macaya (San José, Costa Rica) and G. Matassi (Nottingham, U.K.) for discussions. This research was supported by a postdoctoral fellowship from Association Universitaire d'Expression Partiellement et de Langue Française-Université des Réseaux d'Expression Française to N.C.

1. Bernardi, G., Olofsson, B., Filipski, J., Zerial, M., Salinas, J., Cuny, G., Meunier-Rotival, M. & Rodier, F. (1985) *Science* **228,** 953–958.
2. Bernardi, G. (1989) *Annu. Rev. Genet.* **23,** 637–661.
3. Bernardi, G. (1993) *Gene* **135,** 57–66.
4. Bernardi, G. (1995) *Ann. Rev. Genet.*, in press.
5. Salinas, J., Matassi, G., Montero, L. M. & Bernardi, G. (1988) *Nucleic Acids Res.* **16,** 4269–4285.
6. Matassi, G., Montero, L. M., Salinas, J. & Bernardi, G. (1989) *Nucleic Acids Res.* **17,** 5273–5290.
7. Montero, L. M., Salinas, J., Matassi, G. & Bernardi, G. (1990) *Nucleic Acids Res.* **18,** 1859–1867.
8. Jofuku, K. D. & Goldberg, R. B. (1988) in *Plant Molecular Biology: A Practical Approach*, ed. Shaw, C. H. (IRL, Oxford), pp. 37–66.
9. Bünemann, M. & Dattagupta, N. (1973) *Z. Biochim. Biophys. Acta* **331,** 341–348.
10. Cortadas, J., Macaya, G. & Bernardi, G. (1977) *Eur. J. Biochem.* **76,** 13–19.
11. Church, G. M. & Gilbert, W. (1984) *Proc. Natl. Acad. Sci. USA* **81,** 1991–1995.
12. Matassi, G., Melis, R., Kuo, K. C., Macaya, G., Gehrke, C. W. & Bernardi, G. (1992) *Gene* **122,** 239–245.
13. Sakakibara, H., Watanabe, M., Hase, T. & Sugiyama, T. (1991) *J. Biol. Chem.* **266,** 2028–2035.
14. Winning, B. M., Day, C. D., Sarah, C. J. & Leaver, C. J. (1991) *Plant Mol. Biol.* **17,** 305–307.
15. Rothermel, B. A. & Nelson, T. (1988) *J. Biol. Chem.* **264,** 19587–19592.
16. Colasanti, J., Tyers, M. & Sundaresan, V. (1991) *Proc. Natl. Acad. Sci. USA* **88,** 3377–3381.
17. Matsuoka, M., Ozeki, Y., Yamamoto, N., Hirano, H., Kano-Murakami, Y. & Tanaka, Y. (1988) *J. Biol. Chem.* **263,** 11080–11083.
18. Montoliu, L., Rigau, J. & Puigdomènech, P. (1989) *Plant Mol. Biol.* **14,** 1–15.
19. Worrell, A. C, Bruneau, M., Summerfelt, K., Boersig M. & Voelker, T. A. (1991) *Plant Cell* **3,** 1121–1130.
20. Chandler, V. L., Radicella, J. P., Robbins, T. P., Chen, J. & Turks, D. (1989) *Plant Cell* **1,** 1175–1183.
21. Veit, B., Volbrecht, E., Mathern, J. & Hake, S. (1990) *Genetics* **125,** 623–631.
22. Sullivan, T. D., Strelow, L. I., Illingworth, C. A., Phillips, R. L. & Nelson, O. E. (1991) *Plant Cell* **3,** 1337–1348.
23. Viret, J.-F., Schantz, M.-L. & Schantz, R. (1990) *Nucleic Acids Res.* **18,** 7179.
24. Quigley, F., Martin, W. F. & Cerff, R. (1988) *Proc. Natl. Acad. Sci. USA* **85,** 2672–2676.
25. Pintor-Toro, J. A., Langridge, P. & Feix, G. (1982) *Nucleic Acids Res.* **10,** 3845–3860.
26. Matassi, G., Melis, R., Macaya, G. & Bernardi, G. (1991) *Nucleic Acids Res.* **19,** 5561–5567.
27. Flavell, R. B., Bennett, M. D., Smith, J. B. & Smith, D. B. (1974) *Biochem. Gene.* **12,** 257–269.
28. Hake, S. & Walbot, V. (1980) *Chromosoma* **79,** 251–270.
29. Springer, P. S., Edwards, K. J. & Bennetzen, J. L. (1994) *Proc. Natl. Acad. Sci. USA* **91,** 863–867.
30. Capel, J., Montero, L. M., Martinez-Zapater, J. M. & Salinas, J. (1993) *Nucleic Acids Res.* **21,** 2369–2373.
31. Zoubak, S., Richardson, J. H., Rynditch, A., Höllsberg, P., Hafler, D. A., Boeri, E., Lever, A. M. L. & Bernardi, G. (1994) *Gene* **143,** 155–163.
32. Bennetzen, J. L. & Freeling, M. (1993) *Trends Genet.* **9,** 259–261.
33. Ahn, S. & Tanksley, S. D. (1993) *Proc. Natl. Acad. Sci. USA* **90,** 7980–7984.
34. Flavell, R. B., Gale, M. D., O'Dell, M., Murphy, G. & Moore, G. (1993) *Chromosomes Today* **11,** 199–213.
35. Gill, K. S., Gill, B. S. & Endo, T. R. (1993) *Chromosoma* **102,** 374–381.
36. Moore, G., Abbo, S., Cheung, W., Foote, T., Gale, M., Koebner, R., Leitch, A., Leitch, I., Money, T., Stancombe, P., Yano, M. & Flavell, R. (1993) *Genomics* **15,** 472–482.
37. Antequera, F. & Bird, A. P. (1988) *EMBO J.* **7,** 2295–2299.
38. Gardiner-Garden, M. & Frommer, M. (1992) *J. Mol. Evol.* **34,** 231–245.
39. Gardiner-Garden, M., Sved, J. A. & Frommer, M. (1992) *J. Mol. Evol.* **34,** 219–230.
40. Saccone, S., De Sario, A., Della Valle, G. & Bernardi, G. (1992) *Proc. Natl. Acad. Sci. USA* **89,** 4913–4917.
41. Saccone, S., De Sario, A., Wiegant, J., Raap, A. K., Della Valle, G. & Bernardi, G. (1993) *Proc. Natl. Acad. Sci. USA* **90,** 11929–11933.
42. Gyapay, G., Morissette, J., Vignal, A., Dib, C., Fizames, C., Millasseau, P., Marc, S., Bernardi, G., Lathrop, M. & Weissenbach, J. (1994) *Nat. Genet.* **7,** 246–339.